Email Analytics

Email Analytics System allows us to understand the email traffic by date, hour and year as well as graph analysis. In this project, we'll learn how we can use Apache Sqoop, Apache HDFS, Apache Oozie, and Apache PIG to design an end-to-end data pipeline that will enable us to analyze Email data. Following tasks need to be accomplished in order:

MileStone1: Download the Enron mysql dump(email data) from following link and import the data into mysql database.

https://www.cs.purdue.edu/homes/jpfeiff/enron-mysqldump.tar.gz

MileStone2: Analyze the dataset in mysql tables to get overview of imported data. Check the appendix at the end of this document that provides information about fields.

MileStone3: Using HUE, dry run an oozie workflow with sample actions for each type: hdfs, shell, pig, hive, marpreduce and sqoop.

MileStone4: With JobDesigner in HUE, create the sqoop action to import data from mysql tables into HDFS.

MileStone5: With JobDesigner in HUE, create the pig action to perform the following data preparation steps on HDFS data:

a) Load the Message and RecipientInfo data

b) Join the above datasets to get unified Email data with following fields: message_id,

date, from, to, ccs, bccs, subject, body

c) The date must be transformed from database custom format to ISO format.

d) Transform the data into AVRO format

MileStone6: Using JobDesigner in HUE, develop PIG/HIVE/MR jobs to develop following basic statistics from AVRO enron data:

- **Mails per day of the week**: From Monday-Sunday, mails on each day
- **Mails per hour of the day**: From 0 hours to 23 hours, mails by hour
- **Mails per month**: Emails received by month across all years
- **Mails per year:** Emails received across all year

- **Mails received**: Top 50 people who received more email
- **Mails sent**: Top 50 people who sent more email

MileStone7: Create the tables in mysql and hbase for storing the email statistics.

MileStone8: Using JobDesigner in HUE, create the sqoop action to export the results from HDFS/HIVE tables to mysql tables.

MileStone9: Using JobDesigner in HUE, create the pig action to export the results from HDFS/HIVE tables to hbase tables.

MileStone10: Using JobDesigner in HUE, create the pig action to export the results from HDFS/HIVE tables to hbase tables.

Appendix

Following Section provides the detils of the tables available in mysql dump and their field description.

Employeelist

- eid: Employee-ID
- firstName: First name
- lastName: Last name
- Email_id: Email address (primary). This one can be found in the other tables/dataframes and is useful for matching.
- Email2: Additional email address that was replace by the primary one.
- Email3: See above
- Email4: See above
- folder: The user's folder in the original data dump.
- status: Last position of the employee. "N/A" are unknown.

Message

- mid: Message-ID. Refers to the rows in recipientinfo and referenceinfo.
- sender: Email address (updated)
- date: Date.
- message_id: Internal message-ID from the mailserver.

- subject: Email subject
- body: Email body.
- folder: Exact folder of the e-mail inclusing subfolders.

Recipientinfo

(Note: If an email is sent to multiple recipients, there is a new row for every recipient!)

- rid: Reference-ID
- mid: Message-ID from the message-table/-dataframe
- rtype: Shows if the reciever got the email normally (“to”), as a carbon copy (“cc”)

  or a blind carbon copy (“bcc”).

- rvalue: The recipient’s email address.

Referenceinfo

- rfid: referenceinfo-ID
- mid: Message-ID
- reference: Contains the whole email with shortend headers.